УДК 025.32
ББК 78.52

БВ
Ресурсы
и технологии

Сбор, обработка и хранение библиографических записей с использованием технологий семантической паутины

Реализация проекта по семантической интеграции библиографических записей позволила решить актуальные задачи: разработана онтология предметной области, созданы модули взаимодействия с различными автоматизированными библиотечными информационными системами, библиографические записи сконвертированы из различных форматов в RDF, обогащены за счет информации, полученной из разных источников, и опубликованы в соответствии с принципами Linked Open Data. Оперирование одним из самых крупных в мире массивов библиографических записей потребовало использовать узкоспециализированные протоколы доступа к информации, высокопроизводительные алгоритмы обработки и масштабируемые решения хранения данных.

Ключевые слова: *связанные данные, библиографическая запись, выявление дублетных записей.*

Олег Николаевич Шорин,
заместитель генерального директора по информатизации Российской национальной библиотеки

В Российской государственной библиотеке (РГБ) и Российской национальной библиотеке (РНБ) реализуется проект по публикации библиографических записей с использованием технологий семантической паутины. Основная цель проекта — создание программной системы, которая в автоматическом режиме собирала бы библиографические записи из различных библиотек, осуществляла связывание данных и публиковала их в соответствии с принципами Linked Open Data (LOD, Связанные открытые данные). Наличие открытого доступа к крупнейшему массиву данных, ориентированному на использование не только человеком, но и автоматизированными средствами, позволит создать новые высокоинтеллектуальные онлайн-сервисы, которые окажут значительное влияние на развитие культуры и книжной отрасли.

Для создания модульной системы публикации данных, способной без значительных усилий подключать новых участников, работы были разбиты на несколько этапов [4]:

- разработка онтологии предметной области на базе существующих решений;
- осуществление интеграции с автоматизированными библиотечными информационными системами;
- проведение конвертации библиографических записей из форматов MARC 21 и Rusmarc в унифицированный формат MODS;
- решение вопроса о хранении сконвертированных данных;
- осуществление взаимного обогащения данных из различных библиотек;
- выбор данных для связывания и публикации данных в пространстве LOD;
- реализация модуля визуализации полученного результата.

Данная система имеет распределенную структуру: библиотеки выступают поставщиками библиографических записей, которые аккумулируются, хранятся и обрабатываются на центральном сервере. Нетрудно предположить, что именно центральный сервер является узким звеном в технологическом процессе, поэтому к выбору протоколов взаимодействия с библиотеками, алгоритмов обработки и программных систем хранения данных необходимо подходить с особой тщательностью. Перечисленные компоненты должны обладать такими свойствами, как масштабируемость, интероперабельность, высокая производительность, соответствие общепринятым стандартам.

В мире существует ряд проектов, направленных на интеграцию библиотечных данных и использующих схему аккумуляции данных на центральном сервере из распределенных источников. Среди них можно выделить следующие проекты: Сводный каталог библиотек России (СКБР), Всемирный каталог WorldCat, Европейская цифровая библиотека Europeana. Тщательный анализ используемых в этих проектах решений является залогом успешного функционирования сервиса по семантической интеграции библиографических записей.

Задачи сбора, обработки и хранения библиографических записей

Выполняя свои уставные функции, библиотеки создают библиографические записи на экземпляры, хранящиеся в их фондах. Общее количество записей только в двух крупнейших библиотеках страны — РГБ и РНБ — составляет несколько десятков миллионов. Поиск и получение новых и обновленных записей в таком огромном массиве постоянно меняющейся информации представляет собой отдельную задачу — сбора библиографических записей. Если выбранный способ будет сильно загружать центральный сервер, то при увеличении числа библиотек, участвующих в проекте, работоспособность всей системы в целом не может быть гарантирована.

Другой немаловажной задачей является вопрос выбора способа хранения полученных сведений, поскольку существуют различные механизмы хранения и предоставления доступа к данным, которые имеют как преимущества, так и недостатки. Например, хранение библиографических записей непосредственно в реляционной базе данных с конвертацией их в формат RDF [1] «на лету» существенно экономит дисковое пространство за счет отсутствия дублирования, но при этом резко снижается производительность. Это снижение напрямую зависит как от количества хранимых записей, так и от количества поступающих запросов. Анализ существующих подходов по хранению библиографической информации с последующим предоставлением доступа к ней с использованием протоколов SPARQL и HTTP является проблемой, решение которой направлено на снижение нагрузки на центральный сервер.

Разработка и применение эффективных алгоритмов слияния записей из различных библиотек является третьей задачей. При получении десятков миллионов записей из нескольких библиотек сравнение всех записей для выявления дублетных на одни и те же книги становится неприемлемо затратной задачей. Возникает необходимость создания алгоритма, позволяющего существенным образом сузить множество библиографических записей — потенциальных кандидатов на дублетность. Большинство подобных алгоритмов основано на разбиении всего множества данных на кластеры, внутри которых содержатся подобные друг другу записи. Необходимо проанализировать существующие алгоритмы выявления записей на один и тот же объект с учетом присутствия ошибок и аббревиатур, различий стандартов заполнения полей.

Протокол доступа для сбора библиографических записей

Обмен информацией между библиотеками является устоявшейся практикой. Подобный подход позволяет экономить время и усилия, затрачива-

емые на каталогизацию, поскольку количество одних и тех же экземпляров книг в различных библиотеках велико. Существуют различные протоколы, с помощью которых можно обмениваться библиографической информацией. Самые распространенные и поддерживаемые практически всеми существующими автоматизированными библиотечными информационными системами (АБИС) — протоколы Z39.50 и OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

Протокол Z39.50 изначально был создан в Библиотеке Конгресса США в начале 1980-х годов. Основным его предназначением являлась унификация поиска в различных АБИС, абстрагированная от структуры хранения информации, поисковых языков и правил составления запросов, выходных форматов [5]. До появления этого протокола доступ к АБИС осуществлялся с использованием HTTP — базового протокола навигации по гипертекстовым документам, никоим образом не оптимизированного для работы с библиографической информацией.

Протокол Z39.50 позволяет провести поиск библиографических записей в различных АБИС, используя один и тот же синтаксис. Он широко используется при создании метапоисковых систем (позволяющих произвести поиск одновременно в нескольких разных АБИС и агрегировать полученные результаты). Применение этого протокола позволяет нивелировать различия в моделях хранения, запроса и получения записей в конкретных реализациях.

В конце 1990-х гг. разработчики совместно с библиотекарями сформулировали две основные проблемы, связанные с интероперабельностью цифровых хранилищ данных:

- конечные пользователи вынуждены оперировать с различными поисковыми интерфейсами, предлагаемыми разными системами;
- отсутствие механизмов совместного использования метаданных, которые были бы машинно-ориентированными [10].

Для решения этих проблем предлагалось два альтернативных пути развития: перекрестный поиск информации в различных хранилищах с использованием протокола Z39.50 и аккумуляция данных на центральном сервере из разных источников данных. Подход с использованием перекрестного поиска информации имеет очень большой недостаток: система существенно деградирует при увеличении количества источников, так как скорость работы всей системы равна скорости работы самого медленного звена. В экспериментах, проведенных Networked Computer Science Technical Reference Library (NCSTRL) было показано, что при увеличении количества источников до ста и более система, построенная на перекрестном поиске, переставала быть работоспособной [10].

Было предложено создать новый протокол — OAI-PMH, который был бы предназначен для автоматического сбора метаданных из различных АБИС и аккумуляции их на центральном сервере. Протокол OAI-PMH не является конкурентом Z39.50, поскольку не предназначен для поиска. Основное его назначение — быстро, эффективно, с минимальными затратами собрать библиографические записи вместе, которые впоследствии могут быть использованы для создания других сервисов.

Одним из основных требований протокола OAI-PMH является предоставление поставщиками XML метаданных в формате Dublin Core. Другое немаловажное требование — обязательная реализация запросов, основанных на дате последней модификации библиографических записей.

Поскольку структура системы интеграции библиографических записей представляет собой распределенную конфигурацию с выделенным центральным сервером, на котором агрегируются записи из различных источников, то наиболее подходящим для сбора информации является узкоспециализированный протокол, изначально предназначенный для решения именно этой задачи, — OAI-PMH. Именно он и используется для агрегации библиографических записей из АБИС различных библиотек.

Следует отметить, что WorldCat на базе OAI-PMH создал специальную программу Java-сервлет — OAICat [14], которую держатели метаданных могут адаптировать и установить на своем оборудовании для автоматического сбора центральным сервером WorldCat обновленных записей. Europeana также создала сервис REPOX [13], который основан на использовании OAI-PMH в качестве базового протокола для сбора ресурсов с тысяч различных серверов, расположенных по всей Европе. Нельзя не сказать, что обновленная версия СКБР, известная как СКБР2, агрегирует записи с помощью протокола OAI-PMH [2]. Использование OAI-PMH для аккумуляции библиографических записей из различных источников в таких крупных мировых проектах еще раз убеждает в правильности выбора протокола доступа.

Обработка библиографических записей

Априори известно, что при получении библиографических записей из разных библиотек часто будет возникать ситуация, когда на один и тот же объект будет иметься несколько записей. Они могут отличаться как по формату, так и по полноте заполнения, поскольку в различных учреждениях процессы каталогизации не одинаковы. Например, состав дополнительных элементов, точек доступа может быть разным, системы классификации и предметизации, использование аббревиатур также могут отличаться. Кроме того, запись может содер-

жать ошибки, опечатки, вызванные банальной человеческой оплошностью и невнимательностью.

Вопросы интеграции библиографических записей из различных источников с последующим их объединением и обогащением давно находятся в фокусе внимания ученых. И. Феллеги и А. Сантер, одни из основоположников этого направления, разработали математическую модель, позволяющую разделить множество записей на несколько кластеров [9]. В кластер попадают записи, которые в терминах той или иной метрики располагаются недалеко друг от друга. Для выявления дублетных записей достаточно сравнить записи, входящие в состав одного кластера, что значительно снижает количество сравнений.

Дж. Хилтон развил идеи И. Феллеги и А. Сантера, распространив их не только на проблему выявления дублетных записей, но и создания на их основе обогащенной записи, содержащей объединенную информацию из нескольких записей, с последующим удалением дублетных, содержащих менее полную информацию [11]. Он показал, что процесс разбиения записей на кластеры должен предваряться процедурой нормализации — набором правил, применение которых приводит библиографические записи к некоему единообразному виду. К правилам нормализации можно отнести удаление избыточных пробельных символов, приведение строк к одному регистру, замена общепризнанных аббревиатур и обозначений на унифицированные. Например, записанные в разных форматах даты (Sept. 1987, 09.1987) заменялись на какой-то один формат. В нашем случае процесс нормализации встроен в конвертер, преобразующий записи из форматов MARC 21 и Rusmarc в MODS.

Таким образом, в системе интеграции библиографических записей достаточно создать обогащенную запись, объединяющую несколько описаний на один и тот же объект, и удалить избыточные записи, не содержащих новую информацию.

Поскольку зачастую дополнительная информация, указанная в записи в виде ссылок, оказывается недоступна (например, получить полный текст произведения по ссылке из описания невозможно из-за ограничений, накладываемых авторским правом), то процесс выявления дублетных записей может ориентироваться только на информацию, содержащуюся в самой записи. Поэтому целый класс алгоритмов, использующих дополнительную информацию, не может нами использоваться. Нам доступны только основные поля записи, т. е. по сути можно оперировать лишь строками.

Одними из наиболее распространенных метрик для расчета расстояния между строками являются меры Хемминга, Евклида, Левенштейна, Джаро-Винклера, Рэтклиффа-Обершелпа. Д.Н. Рубцов и В.Б. Барахнин считают, что для простейшего случая, когда сравнение записей осуществляется только по полям «Автор» и «Название», достаточно использование одного из методов динамического программирования, предложенного Хиршбергом [3]. Данный метод обладает высокой эффективностью и относительно простой реализацией. Они описывают также ряд исключений, когда несколько записей при формальной практически полной идентичности содержат информацию о различных объектах. В частности, к таким исключениям можно отнести описания на отдельные тома многотомных изданий.

Таким образом, библиографическая запись, получаемая из АБИС одной из библиотек, проходит через следующую технологическую цепочку:

- преобразование записи в унифицированный формат MODS; поскольку MODS основан на XML, то все последующие преобразования записи можно выполнить с использованием широко распространенных утилит;
- нормализация записи, в процессе которой удаляются лишние пробельные символы, строки приводятся к единому регистру, общепризнанные аббревиатуры и обозначения заменяются унифицированными представлениями;
- сравнение записей, основанное на детерминистическом алгоритме: при совпадении ISBN объекта описания в записи можно утверждать, что записи сделаны для одного и того же объекта;
- в случае отрицательного результата производится вероятностное сравнение записей с использованием алгоритма Хиршберга, при этом учитываются исключения, описанные в [3];

• создание обогащенной записи, содержащей информацию из всех библиографических записей на один и тот же объект.

Основываясь на опыте Консорциума нотного материала библиотек США [8], который объединяет информацию из 31 организации и содержит более 228 тыс. библиографических записей в формате MODS, было принято решение хранить на центральном сервере как обогащенные записи, так и первоначальные. Первоначальные библиографические записи могут пригодиться при поступлении измененной, отредактированной записи из АБИС библиотеки. В этом случае для избежания противоречий проще заново создать обогащенную запись, а не выявлять разницу между первоначальной и измененной записью.

Хранение данных

В 2006 году Т. Бернерс-Ли сформулировал четыре основных принципа связанных данных [6]:

• использование унифицированных идентификаторов ресурса URI (Uniform Resource Identifier) в качестве имен сущностей;

• применение HTTP URI для реализации возможности обращения по именам, для того чтобы они могли быть найдены как людьми, так и программными системами;

• предоставление полезной информации о сущности при обращении по URI, используя стандартизованные форматы;

• включение ссылок на другие связанные URI для облегчения поиска.

Формально получается, что для публикации данных достаточно иметь файловое хранилище и веб-сервер, который при попытке доступа по URI предоставляет необходимую информацию в формате RDF. Очевидно, что такой подход не позволяет использовать SPARQL-точку доступа, наличие которой стало стандартом де-факто.

Однако этот подход плох не только отсутствием SPARQL-точки доступа: при обращении к данным по URI должна происходить автоматическая конвертация файлов данных в формат RDF для каждого запроса. Подобная архитектура не является масштабируемой, поскольку при увеличении объема хранящейся информации и числа запросов производительность системы сильно деградирует.

Для решения перечисленных проблем используют специализированные хранилища RDF-триплетов. Очевидно, что необходимость хранения данных в сконвертированном виде приводит к дополнительному расходу пространства на сервере хранения. Но выгода от использования специализированных хранилищ гораздо больше: применение хранилищ RDF-триплетов позволяет достигнуть необходимого уровня масштабируемости, надежности, безопасности и быстродействия. Наиболее распространенными хранилищами являются 4store и TDB, входящие в состав интегрированной среды Jena.

Опыт использования хранилища 4store в проекте немецких научных библиотек Linking Open Bibliographic Data (LOBID) показал [7], что такое программное решение имеет ограниченную производительность: если запустить поиск на 700 млн RDF-триплетах, полученных из 16 млн библиографических записей, он будет выполняться недопустимо долго. Поскольку система семантической интеграции библиографических записей оперирует сопоставимыми объемами информации, было принято решение использовать интегрированную среду Jena.

Jena является быстроразвивающимся проектом, активно поддерживаемым сообществом программистов. В состав этой среды входят компоненты, которые позволяют:

• хранить данные как в базе данных SQL, так и непосредственно в специализированном хранилище триплетов;

• использовать API для получения RDF-триплетов, информации об используемой онтологии, выполнения SPARQL-запросов;

• с помощью конвертеров сериализовать полученные RDF-триплеты любым из способов: RDF/XML, Turtle, N-Triples, RDF;

• предоставить RDF-данные и ответы на SPARQL-запросы с использованием протокола HTTP.

Очевидно, что при росте количества хранимых триплетов производительность любого специализированного хранилища начнет деградировать. Возможным решением в данной ситуации является использование распределенных файловых систем, позволяющих распределить нагрузку между различными серверами [12].

В процессе создания системы семантической интеграции библиографических записей необходимо было устранить проблемы, связанные со сбором, хранением и обработкой больших массивов. Изначальное решение о создании модульной системы позволило разбить основную задачу на ряд мелких подзадач, для которых можно применить эффективные алгоритмы, использовать высокопроизводительные протоколы взаимодействия и узкоспециализированное программное обеспечение. Основываясь на теоретическом и практическом опыте ведущих зарубежных и отечественных специалистов были приняты решения, которые позволили увеличить масштабируемость, гибкость, отказоустойчивость и интероперабельность всей системы в целом. В частности, для сбора библиографических записей из библиотек используется узкоспециализированный протокол OAI-PMH, непосредственно разработанный для аккумуляции данных из разнородных источников информации.

Для всех записей, собранных из различных источников, применяется ряд преобразований, ко-

торые позволяют привести данные к унифицированному формату, эффективно выявить дублетные записи и создать обогащенную запись, основываясь только на информации из основных полей записи. Цепочка преобразований включает в себя нормализацию записей, сравнение с помощью детерминистического и вероятностного алгоритмов, применение которых позволяет существенным образом уменьшить количество операций сравнения, а также создание обогащенной записи, содержимое которой и публикуется в Linked Open Data.

Для хранения RDF-триплетов была выбрана интегрированная среда Jena, позволяющая предоставлять доступ к RDF-триплетам, сериализовать их несколькими наиболее распространенными способами, выдавать информацию об использованной онтологии, выполнять SPARQL-запросы. При дальнейшем увеличении количества хранимой информации Jena позволяет использовать распределенную файловую сеть для разделения нагрузки между серверами.

Список источников

1. *Жлобинская О.Н.* Как представлять библиографические данные в RDF [Электронный ресурс] / О.Н. Жлобинская. — Режим доступа: http://www.rusmarc.ru/publish/bibdateRDF.pdf
2. *Логинов Б.Р.* СКБР — общероссийский навигатор библиотечных ресурсов и услуг // Ежегодное совещание руководителей федеральных и центральных библиотек субъектов Российской Федерации [Электронный ресурс]. — Режим доступа: http://www.nlr.ru/cms_nlr/vid_news_str.php?id=1637
3. *Рубцов Д.Н.* Выявление дубликатов в разнородных библиографических источниках / Д.Н. Рубцов, В.Б. Барахнин // Вестник НГУ. — Серия: Информ. технологии. — 2009. — Т. 7. — № 3. — С. 86—93.
4. *Серебряков В.А.* Проблемы семантической интеграции библиотечных данных / В.А. Серебряков, О.Н. Шорин // Библиотековедение. — 2014. — № 5. — С. 41—47.
5. *Сысойкина М.А.* Протокол Z39.50 — обоснование необходимости использования [Электронный ресурс] / М.А. Сысойкина. — Режим доступа: http://www.bookresearch.ru/Z3950usage.htm
6. *Berners-Lee T.* Linked Data — Design Issues [Electronic resource] / T. Berners-Lee. — Mode of access: http://www.w3.org/DesignIssues/LinkedData.html
7. *Christoph P.* Building a High Performance Environment for RDF Publishing [Electronic resource] // Semantic Web in Libraries, 2012. — Mode of access: http://swib.org/swib12/slides/Christoph_SWIB12_117.pdf
8. *Davison S.* Enhancing an OAI-PMH Service Using Linked Data : The case of the Sheet Music Consortium [Electronic resource] // Semantic Web in Libraries. — 2013. — Mode of access: http://swib.org/swib13/slides/davison_swib13_123.pdf
9. *Fellegi I.P.* A Theory for Record Linkage / I.P. Fellegi, A.B. Sunter // Journal of the American Statistical Association. — 1969. — Vol. 64. — № 328. — P. 1183—1210.
10. History and Development of OAI-PMH [Electronic resource]. — Mode of access: https://www.oaforum.org/tutorial/english/page2.htm
11. *Hylton J.A.* Identifying and Merging Related Bibliographic Records. / J.A. Hylton. — Cambridge (MA, USA): Massachusetts Institute of Technology, 1996.
12. *Khadilkar V.* Jena-HBase: A Distributed, Scalable and Efficient RDF Triple Store [Electronic resource] / V. Khadilkar [et al.]. — Mode of access: http://www.utdallas.edu/~vvk072000/Research/Jena-HBase-Ext/tech-report.pdf
13. *Pedrosa G.* D5.3.1 — Europeana OAI-PMH Infrastructure — Documentation and final prototype [Electronic resource] / G. Pedrosa [et al.]. — Mode of access: http://www.europeanaconnect.eu/documents/01_Europeana_OAI_PMH_Infrastructure.pdf
14. *Young J.* OAICat [Electronic resource] / J. Young. — Mode of access: http://www.oclc.org/research/activities/oaicat.html?urlm=159694

Контактные данные:
191069, Санкт-Петербург,
ул. Садовая, д. 18;
e-mail: shorin@nlr.ru